Chapter 6
Parallel Computing for Mining Association Rules in Distributed P2P Networks

Huiwei Guan
North Shore Community College, USA

ABSTRACT

Distributed computing and Peer-to-Peer (P2P) systems have emerged as an active research field that combines techniques which cover networks, distributed computing, distributed database, and the various distributed applications. Distributed Computing and P2P systems realize information systems that scale to voluminous information on very large numbers of participating nodes. Data mining on large distributed databases is a very important research area. Recently, most work for mining association rules focused on a single machine or client-server network model. However, this traditional approach does not satisfy the requirements from the large distributed databases and applications in a P2P computing system. Two important challenges are raised, one is how to implement data mining for large distributed databases in P2P computing systems, and the other is how to develop parallel data mining algorithms and tools for the distributed P2P computing systems to improve the efficiency. In this chapter, a parallel association rule mining approach in a P2P computing system is designed and implemented, which satisfies the distribution of the P2P computing system well and makes parallel computing become true. The performance and comparison of the parallel algorithm with the sequential algorithm is analyzed and evaluated, which presents the parallel algorithm features consistent implementation, higher performance, and fine scalable ability.

DOI: 10.4018/978-1-4666-2455-9.ch006

I. INTRODUCTION

Distributed computing deals with hardware and software systems containing multiple processing node or storage element, or multiple programs, running under a loosely or tightly controlled regime. Currently, distributed computing and Peer-to-peer (P2P) systems have emerged as an active research field that combines techniques which cover networks, parallel computing, distributed database, and the various distributed applications. Different from traditional client-server model, a host node in a distributed P2P system has significantly changed the way of information store, sharing, distribution, communication, search and computing. Distributed P2P systems realize information systems that scale to voluminous information on very large numbers of participating nodes (Melucci, 2005; Guan & Chueng, 2000; Wang, 1999; Guan & Li, 1996; Datta, 2005; Guan, 1995; Mehyar, 2005; Guan & Li, 1997; Kamvar, 2005; Guan & Cheung, 1997; Sobolewski, 2006; Guan, 1996; Guan & Sun, 1993; Gao, 2009).

Data mining is a very important research area for database applications, knowledge discovery in databases and data streams, and search engines. The subject can be loosely defined as finding useful patterns or exceptions from a large collection of data. An association rule is an expression $X \Rightarrow Y$, where X is a set of attributes and Y is a single attribute. Intuitively, it means that in the rows of the database if the attributes in X have value "true", Y tends to have value "true" too. The *association rules mining* is to design an efficient algorithm for finding such rules from a database. Recently, most work for mining association rules focused on a single machine or traditional client-server network model. However, this traditional approach does not satisfy the requirements from the large distributed databases and applications in a P2P computing system (Gao, 2009; Kowalczyk, 2003; Agrawal, 1993; Tan, 2006; Hipp, 2000; Guan & Li, 1995; Han, 2006; Guan & Cheung, 1996; Zytkow, 1998). Two important challenges are raised, one is how to implement data mining for large distributed databases in P2P computing systems, and the other is how to develop parallel data mining algorithms and tools for the distributed P2P computing systems to improve the efficiency (Guan & Yu, 2006; Agrawal, 1996; Guan & Li, 1995; Guan & Ip, 2007; Fa, 2006; Guan, 1996; Sun, 2009; Guan & Ip, 1998; Yang, 2008; Guan & Sun, 1992; Kargupta, 2008). In this chapter, a parallel association rule mining approach in a distributed P2P computing system is designed and implemented. First, an overview is given in the section I. Next, a formal specification and some definitions for mining association rules are presented in the section II. Then, a ring P2P computing system architecture is proposed in the section III. Following that, a parallel algorithm for mining association rules on large scale distributed databases in the P2P computing system is designed and implemented in the section IV. An example and analysis are discussed in the section V. The performance and comparison of the parallel algorithm with the sequential algorithm is evaluated in the section VI. Lastly, a summary is addressed in the section VII.

II. FORMAL SPECIFICATION AND DEFINITIONS FOR MINING ASSOCIATION RULES

The formal specification and definitions are presented in this section. Some definitions are given in the section A and the principles of association rules mining are given in the section B.

A. Some Definitions

Definition 1 (itemset): Let D be a database consisting of binary vectors of length n generated by a set of attributes $J = \{J_1, J_2, ..., J_n\}$. An *itemset* $I = \{I_1, I_2, ..., I_m\} \subseteq J$ is a subset of attributes. By definition, any non-empty subset of an *itemset* is also an *itemset*.